Quality Considerations in the CDR Program

Jeff Privette
CDR Program Scientist

30 July 2013

Our 8 Non-Functional Requirements

A Climate Data Record is a time series of measurements of sufficient length, consistency, and continuity to determine climate variability and change (NRC, 2000).

To achieve this, a Climate Data Record shall be:

Accessible	Continuously Assessed/ Improved	Extensible	Preserved	Reproducible	Scientifically defensible	Sustainable	Transparent
Available; Easily obtained, affordable, and understand-able.	Evailuated on a scheduled basis, with the possibility of better methodologies being incorporated	Forward compatible in accommodating new data from existing or new instruments; Able to be adapted for use by others	Secured in perpetuity	Producing consistent results within machine rounding errors	Based on testable hypotheses and methodologies that have been objectively and openly peer-reviewed	Having the potential for long-term maintenance; capable of being continued without exhausting available resources	Openly accountable in every respect

Operational Quality Assurance (QA): A Brief History (Last 365 days)

- Originally not addressed – assumed PI covered it
- Program crafted plan to conduct QA at NCDC
  - PI would provide tools & training to NCDC scientist, who executes
  - Challenges: tool heterogeneity, portability, expertise, etc.
  - Significant time and resources to reach acceptable standard
- Revised & Current Approach (Detailed in Work Agreements)
  - PI delivers a QA Report (2-5 pp.) describing method(s)
  - PI conducts QA; delivers QA results with each update to NCDC
  - Program planning to make these accessible by users

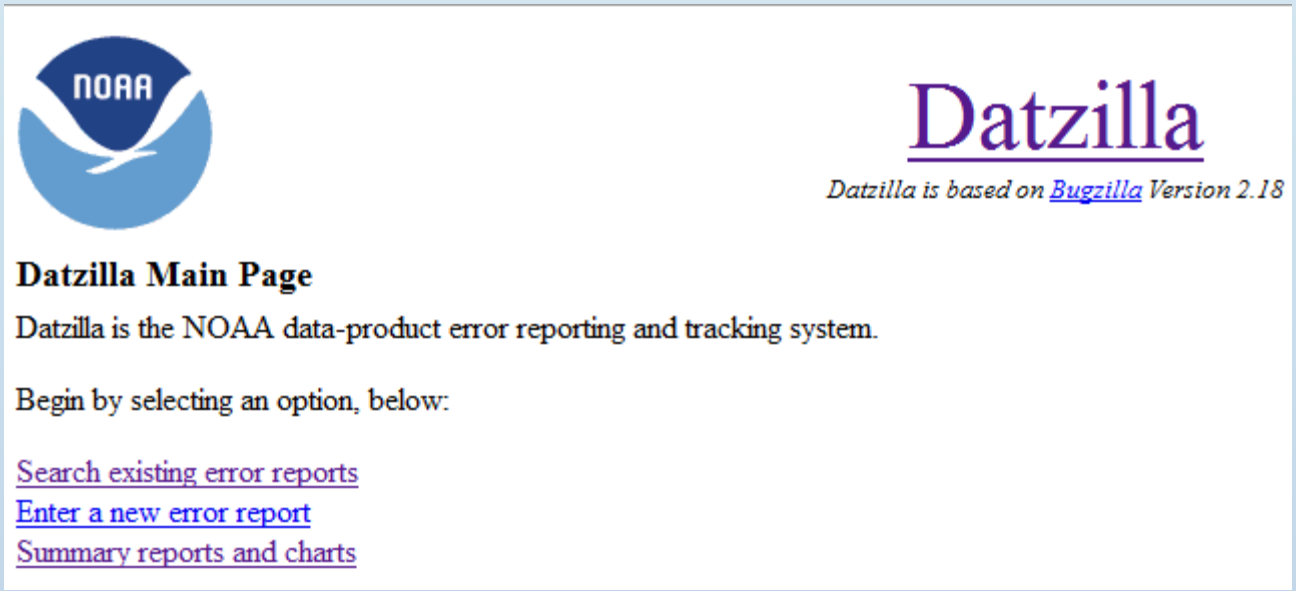
Growing Concern: Data Authenticity

- Once CDR transitioned, the NCDC archived data set is the "gold standard" reference
  - Highly secure, configuration managed, requirements-driven
- Data corruption potential when CDR is served elsewhere
  - Risk to public confidence in trustworthy CDR data sets
- To continue allowing non-NOAA access points, need a solution
  - Digital signatures, Checksum access, etc.

Growing Concern: Product Identification

- Program's requirements impart quality, convey confidence
  - Competitively-selected, sustained, consistent, stable, securely archived, documented, transparent
  - Encourages investment in decision support systems
  - "NOAA CDR" label has value
  - Product identification helps Program stability and longevity
- Labeling should be apparent and consistent at all access points
  - Acknowledgements
  - Links to Archive, Program, etc.
  - Seeking more consistent "look/feel", within reason

Opportunity: Collecting User Feedback and Promoting Community Building



- CDR user forums
  - Existing organizations may offer to host or moderate

International TOVS/ATOVS Working Group

Sub Group for use of TOVS/ATOVS in Climate Studies

Reproducibility Ain't Easy

- All algorithm inputs (flowed, ancillary, Look Up Tables) archived and documented
  - And all of their inputs must be archived…
  - How far back is reasonable, required, affordable?
- Archives must be secure
  - Signed agreements if non-NOAA
- Problem scales with complexity and record length…
  - Transient states may make it virtually impossible
  - Merging in situ and satellite?
- Is archiving test data sets an acceptable compromise?

Long-Term CDR Data Stewardship

Ge Peng
Draft 20130619
ge.peng@noaa.gov

- **Preservation**
  - Product Evaluation
  - Product Acquisition
  - Data Archive
  - Data Governance
- **Accessibility Integrity**
  - Data Search & Discovery
  - Data Availability
  - Data Accessibility
  - Data Usability
  - User Support
  - Security
- **Sustainability Extensibility**
  - Operations & Maintenance
  - Product Update
  - Product Improvement
  - Product Reprocessing
- **Data Quality Defensibility**
  - Data Quality Assessment
  - Data Quality Monitoring
- **Transparency Traceability Reproducibility**
  - Data Provenance
  - Data Reference and Citation
- **Information Content Improvement**
  - Value-Added Products Development

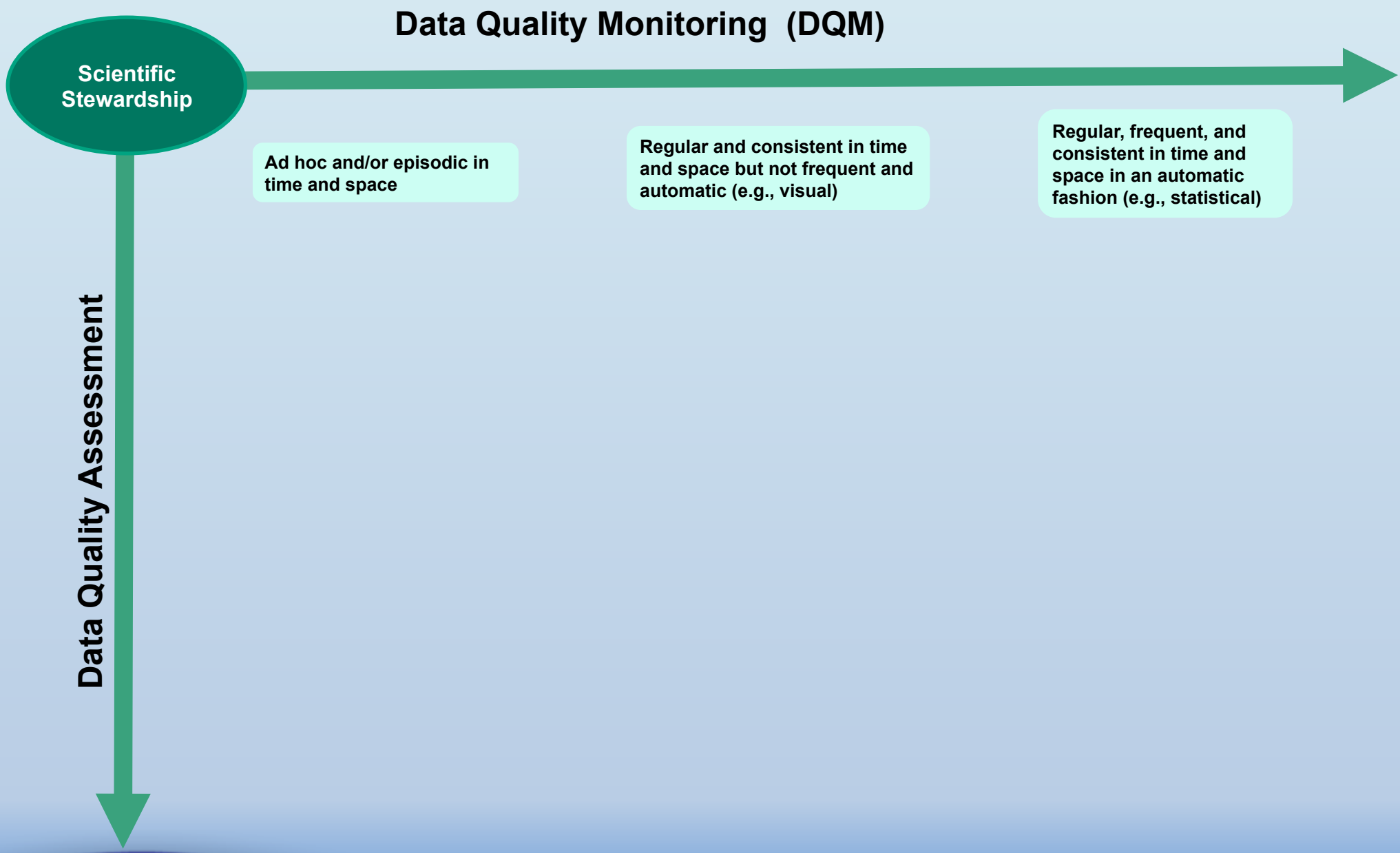
- Configuration Management
- Change Management
- Version Management
- Metadata Management
- Program Management
- Risk Management

- Data provider engagement
- User engagement/feedbacks
- Communication to public
- IT and technical support
- Project support
- Administration & Financial

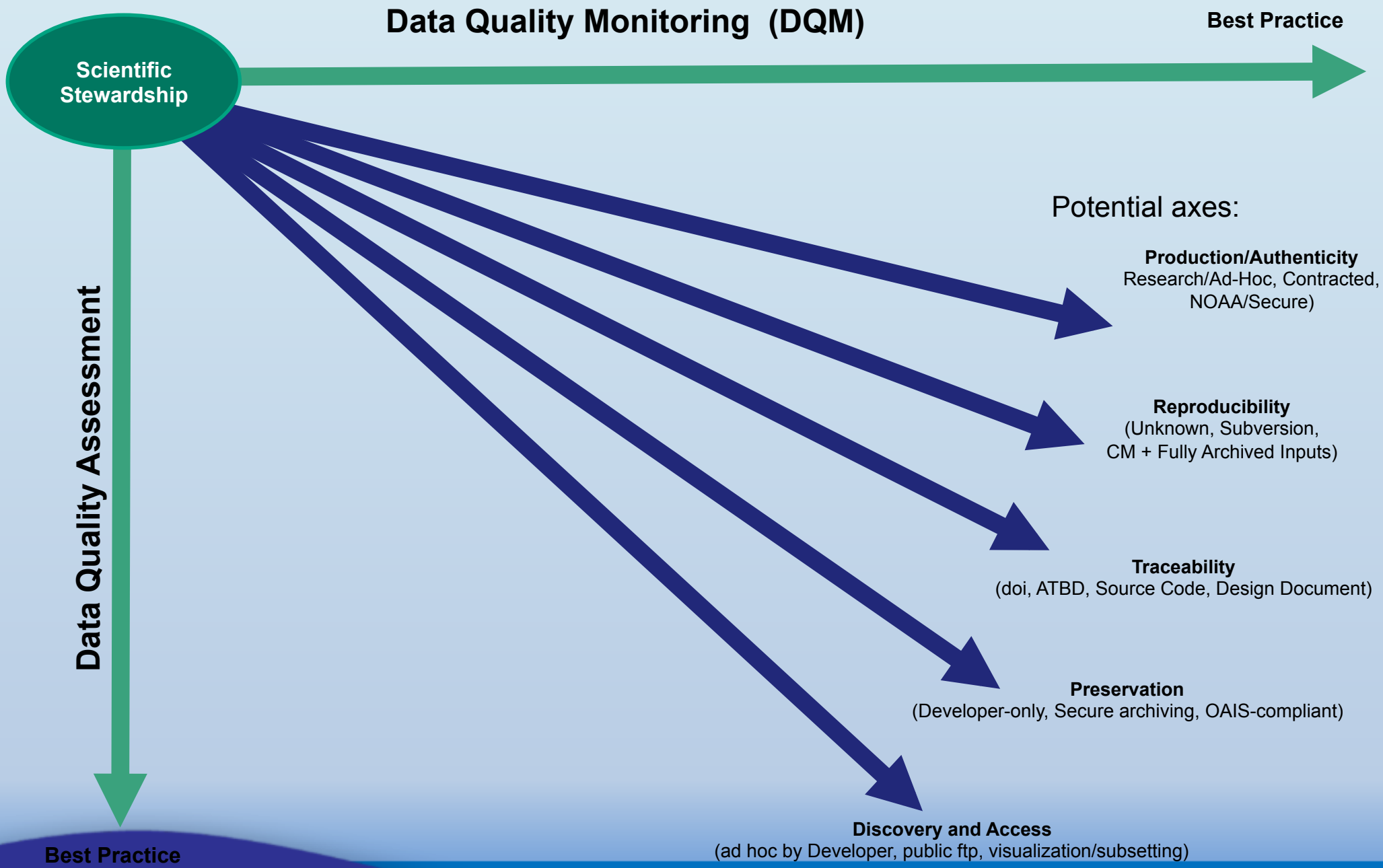
Data Stewardship

- All activities that preserve and improve the information content, accessibility, and usability of data and metadata (NRC, 2007)

Defining Scientific Data Stewardship for NCDC



Defining Scientific Data Stewardship for NCDC



Future Step: Defining Roles & Guidelines

Operations Phase	Algorithm Source and Production	Documentation of Validation and/or Quality Control Process (QC)	Verification of Data Integrity upon Ingest (Checksums)	Sampling & Summarizing of Data Quality (QA)	Monitoring & Flagging of Data Quality (QC)	Algorithm Validation (Scientific & Theoretical Soundness)	Product Validation (Quantitative Uncertainty Estimation)
N/A	Instrument Observation	✓	✓				
IOC	External PI-Generation	✓	✓				
IOC	Internal PI-Generation	✓	✓	✓	✓	✓	✓
FOC	Internal Generation; Externally-Developed Algorithm	✓	✓	✓	✓		
FOC	Internal Generation; Internally-Developed Algorithm	✓	✓	✓	✓		

Summary

- 8 non-functional requirements impart many derived requirements on Quality
- Lack of established standards in community
- Looking at Stewardship framework systematically, but will implement incrementally by building bottom-up